A Stacking-based Hybrid Model with Random Forest as Meta-learner for Diabetes Mellitus Prediction

Aman * and Rajender Singh Chhillar

Department of Computer Science and Applications, M.D. University, Rohtak, India
Email: sei@live.in(a.); chhillar02@gmail.com(R.S.C.)

Manuscript received December 25, 2023; revised May 30, 2024; accepted June 5, 2024; published June 16, 2024.

Abstract—**Diabetes Mellitus (DM) is a condition in which the pancreas is incapable of producing enough insulin for glucose metabolism. Risk factors such as age, hectic schedules, inactivity, patient weight, high blood pressure, and blood sugar level are considered to be the primary cause of type 2 diabetes. Due to misinformation and bad eating habits, the pace of increase in diabetes individuals is problematic. Therefore, a framework employing clinical criteria to diagnose thousands of patients accurately is required. For predicting DM at an early stage based on the risk-based characteristics of a person's health, stacking-based classifier is developed that combines five classifiers, namely Logistic Regression (LR), AdaBoost + Support Vector Machine (SVM), Nave Bayes (NB), Artificial Neural Network (ANN), and k-Nearest Neighbors (k-NN), into a single model and uses Random Forest (RF) as a meta-learner. In addition, the performance of these six classifiers was compared to that of the stacked model using the PIMA Indians Diabetes Database (PIDD) dataset. The outcome of the performance analysis revealed that the proposed model obtained ~85.36% accuracy, which is much higher than the six classifiers.**

Keywords—**AdaBoost, logistic regression, Naïve Bayes, stacking**

I. Introduction

Diabetes is one of the most rapidly expanding chronic illnesses, necessitating an effective predictive model for the diagnosis and prognosis. DM is a metabolic disease in which the body cannot generate sufficient insulin to regulate blood sugar. Diabetes individuals are more prone to have several severe health complications [1]. Every sixth person who has diabetes globally resides in India, making it the country with the second-largest adult diabetes population. The prevalence of diabetes in India has risen by 150 percent in the last three decades [2]. According to the international diabetes organization, this number might reach 134 million by 2045. In the United States, almost one in ten people have diabetes. The majority of instances of type 1 diabetes are caused by an autoimmune illness that targets beta cells that produce insulin and makes the pancreas incapable of making insulin. Type 2 diabetes, which affects 90 to 95% of diabetics, is mainly caused by lifestyle factors, such as being overweight or obese, consuming a diet heavy in fats, sweets, and carbs, and physically inactive. Consequently, the probability of acquiring diabetes increases [3].

After the use of information technologies, the health sector has evolved at a rapid rate. The use of Machine Learning (ML) and Deep Learning algorithms in diabetes research has sparked a flurry of new computational studies, many of which aim to aid physicians in making quick and accurate diagnostic choices [4–7]. To make the best possible adjustments to their daily lives, people with diabetes may now take part in individualized exams of their condition, thanks to the constant improvement of diabetes testing technology. Recent research has classified an accurate rate as superior to current approaches. Early diagnosis of diabetes mellitus is critical. Hence a greater accuracy rate in diabetes prediction is crucial. The team is showcasing many DL and ML approaches to diabetes prediction. There has been a lot of work done on diabetes prediction, but it might be better. Significant health hazards are posed by untreated or delayed diabetes, making this a need. In order to improve prediction performance, this study conducts a comparative analysis of feature selection strategies and data augmentation methods. The key contributions of the paper are summarized as follows: 1) The study proposed a stacking model for the DM prediction, 2) The algorithms ANN, LR, k-NN, NB, and AdaBoost + SVM are employed as the base-learner, while RF serves as the meta-learner, and 3)Accuracy, Precision, Recall, F-measure, Area Under ROC Curve (AUC), and other metrics are evaluated across a wide range of models, including ANN, LR, k-NN, NB, AdaBoost + SVM, and RF, to find the best overall performer.

In Section II, the current work of several scholars on hybrid models for DM prediction is discussed. Section III discusses the methodology for developing stacking-based model for prediction of DM, including 1) the description of the PIDD dataset, 2) the proposed stacking-based model employing RF as a meta-learner, and 3) performance metrics. Finally, section IV compares the performance of the proposed model to that of other models.

II. Related work

Khilwani *et al.* [8] developed a stacked-based model for effective DM prediction. They have used the PIDD dataset from the UCI machine learning repository and implemented using Python. They have stacked six classifiers named SVM, ANN, LR, Decision Tree (DT), Gaussian NB, and RF as base learners. Then passed their prediction to the meta-learner, i.e., Logistic Regression. This model achieved an accuracy of ~82.68%.

Shrestha *et al.* [9] enhanced prediction accuracy and processing speed by proposing a hybridized prediction model. They combined a Support Vector Machine (SVM) + Radial Basis Function (RBF) kernel with a DL approach, namely Long Short-Term Memory Layer (LSTM). Using Python and the PIDD dataset, they have successfully developed the model. A processing time reduction of 3.8 ms and an accuracy of 86.31 percent were the results of their efforts.

Azad *et al.* [10] stressed pre-processing and feature selection phases to improve the prediction model's efficiency. They have proposed a "Prediction Model using Synthetic Minority Oversampling Technique (SMOTE), Genetic

doi: 10.18178/ijml.2024.14.2.1158

Algorithm and Decision Tree" (PMSGD) for DM prediction. This model used SMOTE for handling missing values in the PIDD dataset. Then used genetic algorithm for feature selection. They have finally used a DT for the classifier. Their model achieved an accuracy of ~82.12% and an AUC of ~0.85.

Barik *et al.* [11] proposed a hybrid algorithm for DM prediction. They have combined two boosting algorithms XGBoost with RF. They used a standard PIDD dataset and performed the experience in Python. Their model achieved an accuracy of 74.10%.

Sangien *et al.* [12] surveyed and analyzed the three most common classifiers, i.e., SVM, LR, and RF, against the PIDD dataset. They used 10-cross-fold validation to divide the dataset and deployed their solution on the Waikato Environment for Knowledge Analysis (WEKA). According to their analysis, SVM is the most effective method, with an accuracy of 80%.

Paul & Karn [13] proposed an ANN-based approach to predict DM. They used scaled conjugate gradient backpropagation to minimize the error rate. Their model used the PIDD dataset from the Kaggle repository and implemented it in Python. Their work achieved an accuracy of 77% for finding the presence of DM or not.

Khan *et al.* [14] suggested two stacked-based classifiers to enhance the accuracy of detecting cardiovascular and diabetes-related diseases. The regular UCI repository-obtained dataset has been used and partitioned into 70% training data and 30% test data. They have combined NB, k-NN, LDA, and DT as their fundamental learning algorithms. They implemented the RF as a meta-classifier for cardiovascular disease prediction and found that it increased accuracy to ~88.7%. With SVM as a meta-classifier, they were able to predict DM with an accuracy of 76.46%.

III. Methodology

This section includes a comprehensive description of the proposed stacking-based model, dataset, classifiers, and performance metrics.

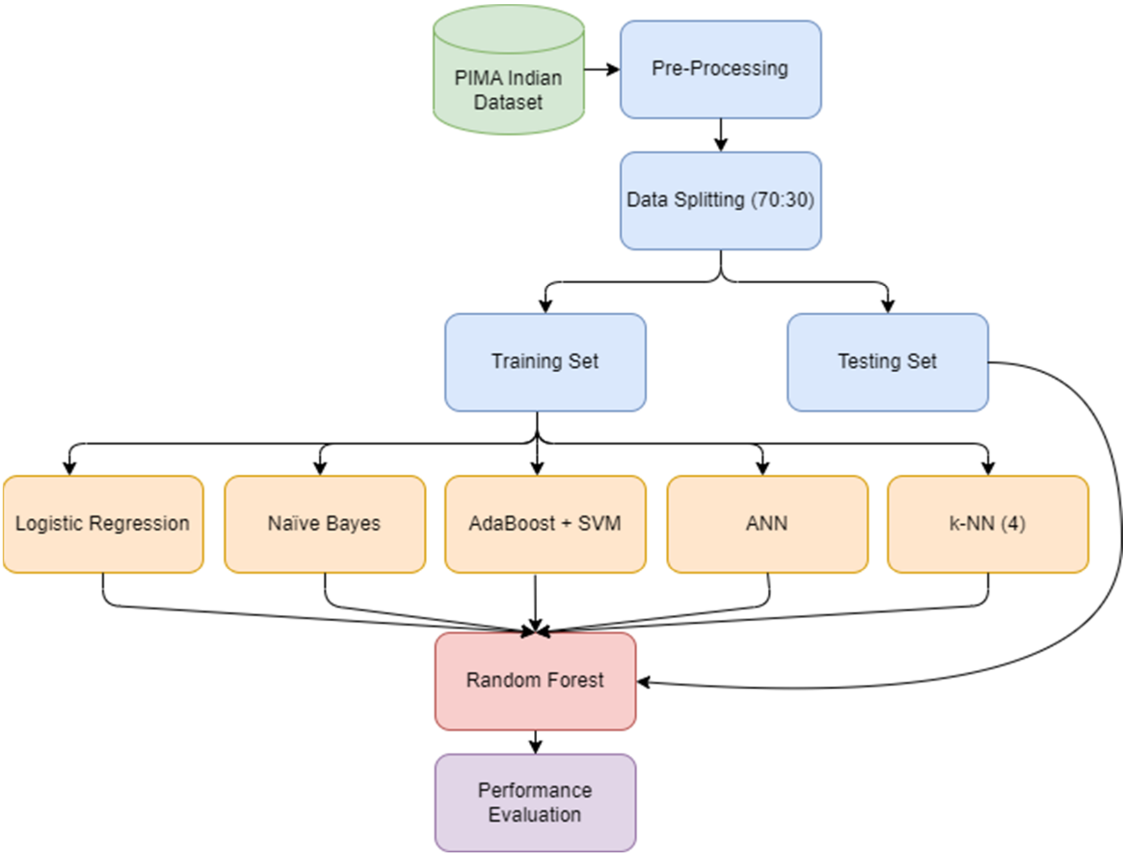

Fig. 1. Workflow of the proposed model.

A. Proposed Stacking algorithm

Stacking is a form of ensemble learning that enables a meta-learner to combine many classifiers. A meta-learner (Level 1) is used to categorize the output produced by several base-learners (Level 0). In order to improve performance in a meta-learning system, any classifier can be employed. The meta-learner, which is responsible for making the final classification, is trained using the aggregated findings of the basic learners.

Fig. 1. demonstrates the process through which our stacking-based paradigm is implemented. There are three phases necessary for applying this proposed paradigm. In this step, the initial training dataset is created and trained using ANN, LR, AdaBoost + SVM, NB, and k-NN. The 70% training dataset is utilized to train ANN, LR, AdaBoost + SVM, NB, and k-NN. After training the five models in the first stage, each model's predictions are obtained. Using the predictions from the level-0 base learner, a new dataset is produced in the third step. The first stage of the base five classifiers will result in a five-dimensional new dataset. A Level-1 classifier known as a meta-learner is used on the dataset during the initial step of the process. In this work, the RF meta-classifier was utilized. In addition, each model will be trained and analysed separately as part of this research in order to evaluate the efficiency and accuracy of the suggested stacking model. A comparison is made between the performance of the proposed stacking model with that of individual classifiers such as ANN, LR, AdaBoost + SVM, NB, k-NN, and RF in terms of Accuracy, Recall, Precision, F-Measure, and Area Under the Curve (AUC). The Proposed Stacking Model for DM Prediction is described in Algorithm 1.

Algorithm 1. The proposed stacking model for diabetes mellitus

```
Input: Diabetes dataset $\mathcal{P} = \{x_i, y_i\}_{i=1}^{n}$
Output: Stacked model $\mathcal{S}$ predictions
begin
Set base learners and meta learner according to $\mathcal{P}$
Step 1. Train all base learners
for $l := 1$ to $L$ do
    Learn a base learner $s_l$ based on $\mathcal{P}$
end for
Step 2. Generate new datasets from $\mathcal{P}$
for $m := 1$ to $n$ do
    Generate a new dataset containing $\{\hat{x}_i, y_i\}$,
    where $\hat{x}_i = \{s_1(x_i), s_2(x_i), \dots, s_L(x_i)\}$
end for
Step 3. Learn random forest as meta learner
Learn a new model $\hat{s}$ based on the newly generated dataset
return $\mathcal{S}(x) = \hat{s}(s_1(x_i), s_2(x_i), \dots, s_L(x_i))$
end
```

B. Dataset

The PIDD dataset, which can be found in the WEKA software's repository [15] and is freely accessible on Kaggle [16], was used in this research. There are diagnostic parameters in the dataset that the data set hopes can be used to determine whether or not a patient has DM.

Table 1. Schema of PIDD dataset

Name	Type	Data Type	Values
class	Dependent	Nominal	tested_positive, tested_negative
insu	Independent	Numerical	0-846
skin	Independent	Numerical	0-99
preg	Independent	Numerical	0-77
plas	Independent	Numerical	0-199
age	Independent	Numerical	21-81
pres	Independent	Numerical	0-122
pedi	Independent	Numerical	0-2.45
mass	Independent	Numerical	0-67

A number of restrictions governed the gathering of such occurrences from a larger database. Every patient whose details were collected is at least 21 years old. Table 1 shows that there are a total of nine attributes in this data collection, one of which being a dependent variable and the others being independent variables. This dataset comprises 768 instances. The dependent variable is unbalanced, and the dataset is prone to missing values, both of which will impact the accuracy of the model. The issue will be addressed while preparing the data.

C. Classifiers

The characteristics of datasets have a significant role in classifier/algorithm selection for stacking. The following algorithms were selected for this study after a thorough evaluation of the relevant literature:

Random Forest is a bagging approach that employs DTs in a parallel manner [17]. Once each DT has been fed training data, many votes may be predicted with high accuracy. When it comes to DT, overfitting is a prevalent problem that may be fixed using RF.

An Artificial Neural Network is a three-layered machine learning classifier in which each layer feeds its output to the one underneath it. The outcomes of the nodes in the Input Layer are sent to the nodes in the Hidden Layer below. Changing the ANN's hidden node count might potentially boost its performance. The output from the Hidden layer is transferred to the output layer at a later time. The main drawback of ANN is that it cannot justify its own actions [18].

Logistic Regression is a method of regression analysis that utilizes the sigmoid function and is developed from Linear Regression. Scaling the y-value from a wide range to an exact interval is accomplished by the use of a sigmoid function in the logistic function (0, 1) [19].

To improve the performance of unreliable binary classifiers (such as DT), researchers have developed the ensemble learning technique known as adaptive boosting (AdaBoost). Here, weak classifiers are introduced incrementally, rather than all at once, like in RF. The number of decision stumps generated is equal to the number of feature variables in the dataset. At first, all of the information available to the various decision trees was given the same value. The model with the lowest Entropy will serve as the starting point for the selection process. Then, a normalized new weight is assigned to each observation depending on the performance and overall error. After that, a new decision stump will be chosen at random with its weight levelled [20].

Naïve Bayes classifier is based on Bayes' theorem. With no effort and no need for a complex iteration model, you may get an iterative parameter estimate that performs admirably even on the biggest datasets. In many cases, the Naive Bayesian classifier outperforms more sophisticated classification techniques despite its apparent lack of complexity [21].

k-Nearest Neighbors is a classification model that organizes data points based on their nearest neighbours. The steps involved in implementing k-NN are rather simple. Converting the data points to vectors at the outset. Next, we use a mathematical computation, such as the Euclidian Equation, to determine how far apart two vector points are; this yields the Manhattan distance. After that, we calculate the probability that these points are analogous to the test data. The most likely vector point is then selected [22].

D. Performance Metrics

The Receiver Operating Characteristic (ROC) Curve is a performance assessment measure for binary classification model(s). Differentiating signals from noise is achieved by following the TPR (True positive rate) to different thresholds in relation to the FPR (False positive rate). The Area Under the Curve is a metric used to evaluate the performance of a classifier. A higher AUC indicates that the model does a better job of separating the two classes.

Accuracy indicates how frequently the model is accurate. Mathematically, it is demonstrated by (1). Precision is the proportion of True Positives ($\mathcal{TP}$) to total positives. Thus, recall informs us, for all diabetic patients, how many accurately recognize as diabetic. The recall is the proportion of True Positives accurately identified by our model. Thus, recall informs us, for all diabetic patients, how many accurately recognize as diabetic. Mathematically, (2) and (3) represent Precision and Recall, respectively. In (4), F-measure is the harmonic mean of Precision and Recall. A researcher may prioritize a high F-Measure rather than trying to strike a balance between Precision and Recall.

$$Accuracy = \frac{(\mathcal{TP} + \mathcal{TN})}{(\mathcal{TP} + \mathcal{FP} + \mathcal{FN} + \mathcal{FN})} \tag{1}$$

$$Precision = \frac{(\mathcal{TP})}{(\mathcal{TP} + \mathcal{FP})} \tag{2}$$

$$Recall = \frac{(\mathcal{TP})}{(\mathcal{TP} + \mathcal{FN})} \tag{3}$$

$$F - Measure = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{4}$$

IV. Results and Discussion

The aim of this work is to ascertain whether or not the patient suffers from DM. Specifically, a 16 GB RAM, 3.2 GHz Intel Core i5 CPU Waikato Environment for Knowledge Analysis (WEKA) machine was used to conduct the research. Following data preparation, the mean value is substituted for missing instances. This will lessen the trade-off between precision and recall. Following the treatment of Missing values, Data records are separated into training sets consisting of 70 percent and test sets consisting of 30 percent. Data mining approaches and algorithms include RF, ANN, and LR. These classifiers get Training data and are then verified using Testing data. Using the WEKA software, the performance of these models was then tested.

The accuracy of RF, K-NN, ANN, NB, AdaBoost + SVM, LR, and the recommended stacking model are compared in Table 2 and Fig. 2. Table 2 illustrates that the RF model has an accuracy of 72.52 percent while the k-NN model has an accuracy of 75.21 percent. 80.43 percent accuracy is a great performance by LR. In addition, ANN possesses an accuracy score of 76.95 percent, whereas NB possesses an accuracy score of 77.82 percent and AdaBoost + SVM scores an accuracy score of ~79.13 percent. The suggested stacking-based model scored an accuracy of ~85.36 percent. This research indicates that coupled classifiers outperform their individual equivalents.

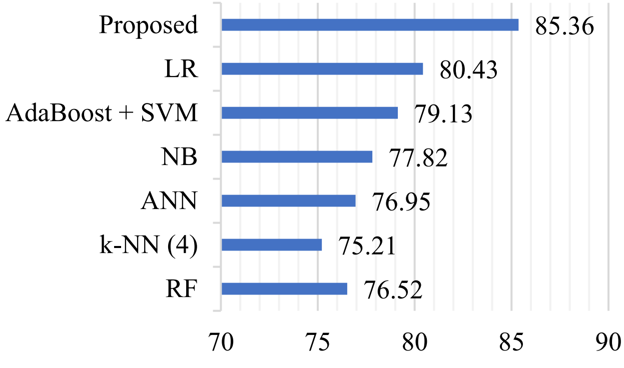

Fig. 2. Comparison of Accuracy (in %) against PIDD dataset.

Table 2. Performance comparison of the models against the PIDD dataset

Model	Accuracy (%)	Precision	Recall	F-Measure	AUC
RF	76.5217	0.756	0.765	0.758	0.832
k-NN (4)	75.2174	0.740	0.752	0.731	0.775
ANN	76.9565	0.784	0.770	0.774	0.792
NB	77.8261	0.770	0.778	0.770	0.830
AdaBoost + SVM	79.1304	0.784	0.791	0.782	0.763
LR	80.4348	0.799	0.804	0.799	0.848
Proposed	85.3659	0.864	0.864	0.850	0.950

In general, models with a ROC Area value near 1 are considered to be effective against the dataset. The suggested stacked-based model on the PIDD dataset had a ROC Area of 0.95. (See Table 2). After the suggested model, LR scored 0.84, making it the second-most viable option for DM prediction.

V. Conclusion

This effort aims to develop and deploy a stacked-based model for the accurate prediction of DM. This model made use of ANN, LR, AdaBoost + SVM, NB, and k-NN as the Base learner and RF as the Meta learner. The experiment is conducted using the WEKA programme and the PIDD dataset. In this study, the suggested model is tested based on precision, recall, accuracy, ROC AUC score, F Measure, and MAE. The accuracy of the proposed model was 85.36%, which was greater than the accuracy of the base learners. The dataset primarily determines the limitations of an algorithm(s) analysis. Since this dataset only contains statistics pertaining to girls under the age of 21. This study may be improved by incorporating additional datasets with both genders and varied ages. Future research might improve the performance of the prediction model by developing and implementing hybrid classifiers based on Deep learning and Metaheuristic algorithms.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Aman collected the data, performed the experiment, analyzed the results, and wrote the manuscript. Rajender Singh Chhillar contributed to the data analysis and assisted in the preparation and revision of the manuscript. All authors have reviewed and approved the final version of the manuscript.

References

[1] A. Darolia and R. S. Chhillar, “Analyzing three predictive algorithms for diabetes mellitus against the Pima Indians dataset,” *ECS Trans.*, vol. 107, no. 1, p. 2697, Apr. 2022, doi: 10.1149/10701.2697ecst.

[2] R. M. Anjana *et al.*, “Macronutrient recommendations for remission and prevention of diabetes in Asian Indians based on a data-driven optimization model: The ICMR-INDIAB national study,” *Diabetes Care*, p. dc220627, Aug. 2022, doi: 10.2337/dc22-0627.

[3] Annual Reports | American Diabetes Association. (Nov. 15, 2022). [Online]. Available: https://diabetes.org/about-us/reports

[4] Aman and R. S. Chhillar, “Analyzing predictive algorithms in data mining for cardiovascular disease using WEKA tool,” *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 12, no. 8, Art. no. 8, 31 2021, doi: 10.14569/IJACSA.2021.0120817.

[5] Aman and R. S. Chhillar, “Disease predictive models for healthcare by using data mining techniques: State of the art,” *Int. J. Eng. Trends*

Technol. - IJETT, vol. 68, no. 10, pp. 52–57, 2020, doi: https://doi.org/10.14445/22315381/IJETT-V68I10P209.

[6] A. Darolia and R. Chhillar, “Analyzing predictive algorithms in data mining for cardiovascular disease using WEKA Tool,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, p. 2021, Jan. 2021, doi: 10.14569/IJACSA.2021.0120817.

[7] A. Darolia and R. Chhillar, “Optimized stacking ensemble for early-stage diabetes mellitus prediction,” *Int. J. Electr. Comput. Eng. IJECE*, vol. 13, pp. 7048–7055, Dec. 2023, doi: 10.11591/ijece.v13i6.pp7048-7055.

[8] V. O. Khilwani, V. Gondaliya, S. Patel, J. Hemnani, B. Gandhi, and S. K. Bharti, “Diabetes Prediction, using Stacking Classifier,” in *Proc. 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, Sep. 2021, pp. 1–6. doi: 10.1109/AIMV53313.2021.9670920.

[9] M. Shrestha *et al.*, “A novel solution of deep learning for enhanced support vector machine for predicting the onset of type 2 diabetes,” *Multimed. Tools Appl.*, Aug. 2022, doi: 10.1007/s11042-022-13582-9.

[10] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, “Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus,” *Multimed. Syst.*, vol. 28, no. 4, pp. 1289–1307, Aug. 2022, doi: 10.1007/s00530-021-00817-2.

[11] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, “Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques,” in *Intelligent and Cloud Computing*, D. Mishra, R. Buyya, P. Mohapatra, and S. Patnaik, eds., Singapore: Springer, 2021, pp. 399–409. doi: 10.1007/978-981-15-6202-0_41.

[12] T. Sangien, T. Bhat, and M. S. Khan, “Diabetes Disease Prediction Using Classification Algorithms,” in *Internet of Things and Its Applications*, K. Dahal, D. Giri, S. Neogy, S. Dutta, and S. Kumar, eds., Singapore: Springer Nature, 2022, pp. 185–197. doi: 10.1007/978-981-16-7637-6_17.

[13] B. Paul and B. Karn, “Diabetes mellitus prediction using hybrid artificial neural network,” in *Proc. 2021 IEEE Bombay Section Signature Conference (IBSSC)*, Nov. 2021, pp. 1–5. doi: 10.1109/IBSSC53889.2021.9673397.

[14] A. Khan, A. Khan, M. M. Khan, K. Farid, M. M. Alam, and M. B. M. Su’ud, “Cardiovascular and diabetes diseases classification using ensemble stacking classifiers with SVM as a meta classifier,” *Diagnostics*, vol. 12, no. 11, Art. no. 11, Nov. 2022, doi: 10.3390/diagnostics12112595.

[15] Datasets - Weka Wiki. (Nov. 15, 2022). [Online]. Available: https://waikato.github.io/weka-wiki/datasets/

[16] *Datasets | Kaggle*. (Dec. 10, 2019). [Online]. Available: https://www.kaggle.com/datasets

[17] G. Biau and E. Scornet, “A random forest guided tour,” *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.

[18] J. J. Hopfield, “Artificial neural networks,” *IEEE Circuits Devices Mag.*, vol. 4, no. 5, pp. 3–10, Sep. 1988, doi: 10.1109/101.8118.

[19] R. E. Wright, “Logistic regression,” in *Reading and Understanding Multivariate Statistics*, Washington, DC, US: American Psychological Association, 1995, pp. 217–244.

[20] Explaining AdaBoost | SpringerLink. (Nov. 15, 2022). [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-41136-6_5

[21] F. Zheng and G. I. Webb, “Tree augmented Naive Bayes,” in *Encyclopedia of Machine Learning and Data Mining*, 2019. doi: 10.1007/978-1-4899-7687-1_850.

[22] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “KNN Model-Based Approach in Classification,” in *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, R. Meersman, Z. Tari, and D. C. Schmidt, eds., Berlin, Heidelberg: Springer, 2003, pp. 986–996. doi: 10.1007/978-3-540-39964-3_62.